Chapter 2
M-Estimators and Half-Quadratic Minimization

In robust statistics, there are several types of robust estimators, including M-estimator (maximum likelihood type estimator), L-estimator (linear combinations of order statistics), R-estimator (estimator based on rank transformation) [77], RM estimator (repeated median) [141], and LMS estimator (estimator using the least median of squares) [133]. When information theoretic learning is applied to robust statistics, the Gaussian kernel in entropy plays a role of Welsch M-estimator and can be efficiently optimized by half-quadratic minimization. Hence, in this chapter, we introduce some basic concepts of M-estimation and half-quadratic minimization.

2.1 M-Estimation

M-estimators are defined as the minima of summation of functions of a dataset. The statistical procedure of evaluating an M-estimator is called M-estimation, which is defined as a generalized maximum-likelihood method for the following cost function [99, 172]:

$$\min_{\theta} \sum_j \phi(e_j|\theta), \tag{2.1}$$

where $\phi(.)$ is differentiable and satisfies four conditions (p. 5291, [99]):

$$\begin{aligned}
&\phi(t) \geq 0;\\
&\phi(0) \geq 0;\\
&\phi(t) = \phi(-t);\\
&\phi(t) \geq \phi(\bar{t}) \text{ for } |t| > |\bar{t}|
\end{aligned}$$

R. He et al., *Robust Recognition via Information Theoretic Learning*,
SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-07416-0_2

Table 2.1 A few commonly used M-estimators ($\phi(.)$) and their corresponding weighting functions ($w(.)$). c is a constant [172]

Type	$\phi(t)$	$w(t)$
ℓ_1	$\lvert t\rvert$	$1/\lvert t\rvert$
ℓ_1-ℓ_2	$2(\sqrt{1+t^2/2}-1)$	$1/\sqrt{1+t^2/2}$
ℓ_p	$\lvert t\rvert^c/c$	$\lvert t\rvert^{c-2}$
Fair	$c^2(\lvert t\rvert/c-log(1+\lvert t\rvert/c))$	$1/(1+\lvert t\rvert/c)$
Huber $\begin{cases} \text{if} \lvert t\rvert \leq c \\ \text{if} \lvert t\rvert > c \end{cases}$	$\begin{cases} t^2/2 \\ c(\lvert t\rvert-c/2) \end{cases}$	$\begin{cases} 1 \\ c/\lvert t\rvert \end{cases}$
Cauchy	$(c^2/2)log(1+(t/c)^2)$	$1/(1+(t/c)^2)$
Geman-McClure	$t^2/(2(1+t^2))$	$1/(1+t^2)^2$
Welsch	$(c^2/2)(1-exp(t/c)^2)$	$exp((t/c)^2)$
Tukey $\begin{cases} \text{if} \lvert t\rvert \leq c \\ \text{if} \lvert t\rvert > c \end{cases}$	$\begin{cases} \frac{c^2}{6}(1-(1-(t/c)^2))^3 \\ c^2/6 \end{cases}$	$\begin{cases} (1-(t/c)^2)^2 \\ 0 \end{cases}$

and θ is a set of adjustable parameters and e_j is an error produced by a learning system [99]. An M-estimator is often solved by the following iteratively reweighted way:

$$\min_{\theta} \sum\nolimits_j w(e_j^t)(e_j^2|\theta), \tag{2.2}$$

where $w(.)$ is a weighting function with respect to $\phi(.)$. The weight $w(e_j^t)$ should be recomputed after each iteration in order to be used in the next iteration. Table 2.1 shows nine M-estimators ($\phi(.)$) and their corresponding weighting functions ($w(.)$).

2.2 Half-Quadratic Minimization

Half-quadratic (HQ) minimization was pioneered in [54, 55] for reconstructing images and signals. Then Idier [80] further studied HQ minimization for image restoration. Champagnat and Idier [21] made a connection between HQ minimization and expectation maximization (EM). And a systematic analysis of the global and local convergence of HQ was given in [3, 112]. Recently, HQ minimization and its minimization functions have been widely used in machine learning and computer vision, such as mean-shift [166], image registration and synthesis [71], and robust feature extraction [165]. It is a general optimization method for convex or non-convex minimization based on conjugate function theory.

Given a differentiable function $f(v) : \mathbb{R}^d \to \mathbb{R}$, the conjugate $f^*(p) : \mathbb{R}^d \to \mathbb{R}$ of the function f is defined as [13]

$$f^*(p) = \max_{v \in dom f} (p^T v - f(v)). \tag{2.3}$$

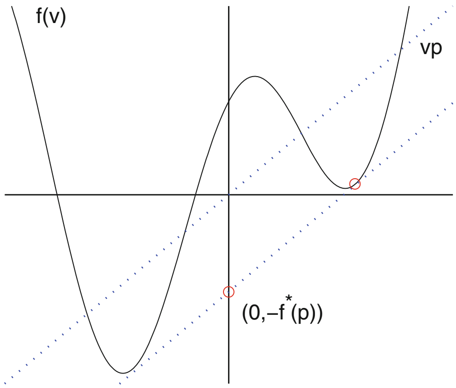


Fig. 2.1 A differentiable function $f(v) : \mathbb{R} \to \mathbb{R}$ and its conjugate function $f^*(p)$ with a value p [13]. As shown by the *dashed line*, the conjugate function $f^*(p)$ is the maximum gap between the linear function vp and $f(v)$, which occurs at a point v where $f'(v) = p$ [13]

The domain of $f^*(p)$ is bounded above on **dom**f [13]. Since $f^*(p)$ is the point-wise supremum of a family of convex functions of p, it is a convex function [13]. If $f(v)$ is convex and closed, the conjugate of its conjugate function is itself, i.e., $f^{**} = f$ [13]. Figure 2.1 gives an illustration of a conjugate function.

Based on conjugate function theory, a loss function in image restoration and signal recovery can be defined as [9, 22, 30]

$$f(v) = \min_p \{Q(v, p) + \varphi(p)\}, \tag{2.4}$$

where $f(.)$ is a potential loss function (such as M-estimators), v is a set of adjustable parameters of a linear system, p is an auxiliary variable in HQ optimization, $Q(v, p)$ is a quadratic function ($Q(v, p) \doteq \sum_j p_j v_j^2$ for $p \in \mathbb{R}^d_+$ and $v \in \mathbb{R}^d$, or $Q(v, p) \doteq ||v - p||_2^2$ for $p \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$), and $\varphi(.)$ is the dual potential function of $f(.)$.[1]

In the two-step iterative shrinkage/thresholding algorithms [161], the minimization function of (2.4) is also known as proximal mapping [9, 30]; in half-quadratic methods, the function $Q(v, p) + \varphi(p)$ is called the resultant (augmented) cost-function of $f(v)$, and can be optimized by a two-step alternating minimization way [22].

2.2.1 *Iterative Minimization*

Let $\phi_v(.)$ be a function on a vector $v \in \mathbb{R}^d$ that is defined as

$$\phi_v(v) \doteq \sum_{j=1}^d \phi(v_j), \tag{2.5}$$

[1]Note that for different types of $Q(v, p)$, the dual potential functions $\varphi(.)$ may be different.

where $\phi(.)$ is a potential loss function in HQ [112,165] and v_j is the jth entry of v. In machine learning and compressed sensing, one often aims to compute the following minimization problem:

$$\min_{v} \phi_v(v) + J(v), \tag{2.6}$$

where $J(v)$ is a convex penalty function on v. According to half-quadratic minimization [54, 55], we know that for a fixed v_j, the following equation holds:

$$\phi(v_j) = \min_{p_j} Q(v_j, p_j) + \varphi(p_j), \tag{2.7}$$

where $\varphi(.)$ is the dual potential function of $\phi(.)$, and $Q(v_j, p_j)$ is the half-quadratic function which can be modeled in the additive or the multiplicative form as shown in Sect. 2.2.2. Let $Q_v(v, p) \doteq \sum_{j=1}^{d} Q(v_j, p_j)$, we have the vector form of (2.7),

$$\phi_v(v) = \min_{p} Q_v(v, p) + \sum\nolimits_{j=1}^{d} \varphi(p_j). \tag{2.8}$$

By substituting (2.8) into (2.6), we obtain that

$$\min_{v}\{\phi_v(v) + J(v)\} = \min_{v,p}\{Q_v(v, p) + \sum_{j=1}^{d} \varphi(p_j) + J(v)\}, \tag{2.9}$$

where p_j is determined by a minimization function $\delta(.)$ that is only related to $\phi(.)$ (See Table 2.2 for specific forms). In HQ optimization, $\delta(.)$ is derived from conjugate function and satisfies that $\{Q(v_j, \delta(v_j)) + \varphi(\delta(v_j))\} \leq \{Q(v_j, p_j) + \varphi(p_j)\}$. Let $\delta_v(v) \doteq [\delta(v_1), \ldots, \delta(v_d)]$, and then one can alternately minimize (2.9) as follows,

$$p^{t+1} = \delta_v(v), \tag{2.10}$$

$$v^{t+1} = \arg\min_{v} Q_v(v, p^{t+1}) + J(v), \tag{2.11}$$

where t indicates the tth iteration. Algorithm 1 summarizes the optimization procedure. At each step, the objective function in (2.9) is employed with respect to a single parameter variable alternatively until it converges.

2.2.2 *The Additive and Multiplicative Forms*

In HQ minimization, the half-quadratic reformulation $Q(v_j, p_j)$ of an original cost-function has two forms [54, 55]: the additive form denoted by $Q_A(v_j, p_j)$ and the multiplicative form denoted by $Q_M(v_j, p_j)$. Specifically, $Q_A(v_j, p_j)$ is formulated as [55],

Algorithm 1 Half-Quadratic Based Algorithms

Input:data matrix X, test sample y, and $v = \mathbf{0}$.
Output:v
while "not converged" **do**
  $p^{t+1} = \delta_v(v)$
  $v^{t+1} = \arg\min_{v} Q_v(v, p^{t+1}) + J(v)$
  $t = t+1$
end while

$$Q_A(v_j, p_j) = (v_j\sqrt{c} - p_j/\sqrt{c})^2, \tag{2.12}$$

where c is a constant and $c > 0$. The additive form indicates that we can expand a function $\phi(.)$ to a combination of quadratic terms and the auxiliary variable p_j is related to v_j. During iterative minimization, the value of v_j is updated and refined by p_j. If a potential function $\phi(.)$ satisfies,

(a) $t \to \phi(t)$ is convex;
(b) $c > 0$ is such that $t \to \{ct^2/2 - \phi(t)\}$ is convex;
(c) $\phi(t) = \phi(-t), \forall v \in r$;
(d) ϕ is continuous on R;
(e) $\lim_{|t|\to\infty} \phi(t)/t^2 < c/2$,

then there is a minimization function $\delta_A(t) = ct - \phi'(t)$ such that [112]

$$\delta_A(t) = \arg\min_{w}(t\sqrt{c} - w/\sqrt{c})^2 + \varphi(w). \tag{2.13}$$

Let $c = 1$, (2.13) takes the form,

$$\delta_A(t) = \arg\min_{w}(t - w)^2 + \varphi(w). \tag{2.14}$$

In the two-step iterative shrinkage/thresholding algorithms [161], the minimization function of (2.13) is also known as proximal mapping [9, 30]. And when $\phi(.)$ is Huber M-estimator, the following additive form holds:

$$\phi_H^{\lambda}(t) = \min_{w}\{(t - w)^2 + \lambda\,|w|\}, \tag{2.15}$$

where absolute function $\lambda\,|w|$ is the dual potential function of Huber M-estimator $\phi_H^{\lambda}(.)$.

The multiplicative form $Q_M(v_j, p_j)$ is formulated in the form [54]

$$Q_M(v_j, p_j) = \frac{1}{2} p_j v_j^2. \tag{2.16}$$

It indicates that we can expand a non-convex (or convex) function $\phi(.)$ to the quadratic term of the multiplicative form. The auxiliary variable p_j is introduced as a data-fidelity term. For v_j, p_j indicates the contribution of v_j to the whole data v. In addition, if a potential function $\phi(.)$ satisfies the following conditions ($t = v_j$ and $w = p_j$),

(a) $\forall t, \phi(t) \geq 0, \phi(0) = 0$;
(b) $\phi(t) = \phi(-t)$;
(c) $\phi(.)$ continuously differentiable;
(d) $\forall t, \phi'(t) \geq 0$;
(e) $\phi'(t)/2t$ continuous and strictly decreasing on $[0, +\infty]$;
(f) $\lim_{t \to +\infty} \{\phi'(t)/2t\} = 0$;
(g) $\lim_{t \to 0^+} \{\phi'(t)/2t\} = M, 0 < M < +\infty$,

then Theorem 2.1 holds.

Theorem 2.1 ([22]). *Let $\phi(.)$ be a potential function that satisfies the above seven conditions, then*

(a) there exists a strictly convex and decreasing dual function $\varphi : (0, M] \to [0, \beta)$, where

$$\beta = \lim_{t \to +\infty} \left(\phi(t) - t^2(\phi'(t)/2t)\right),$$

such that

$$\phi(t) = \inf_{0 < w \leq M} (wt^2 + \varphi(w)),$$

(b) for every fixed $t \geq 0$, the value $\hat{w}$ for which the minimum is reached, i.e., such that

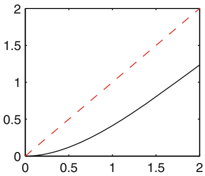
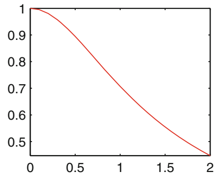
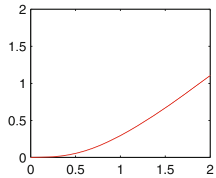
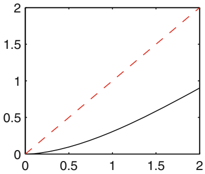
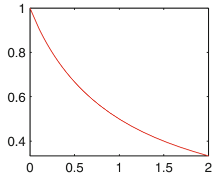
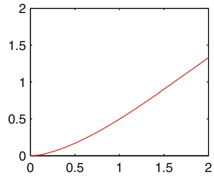
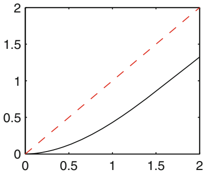
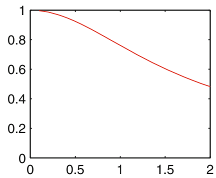
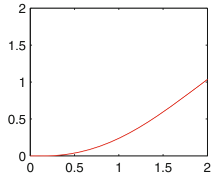
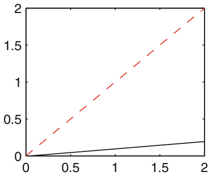
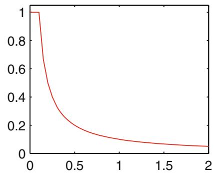
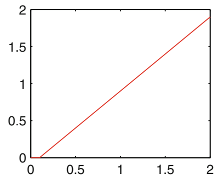
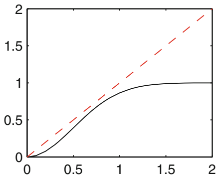
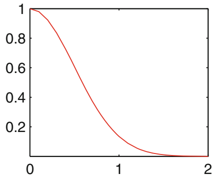
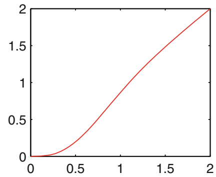
$$\inf_{0 < w \leq M} (wt^2 + \varphi(w)) = (\hat{w}t^2 + \varphi(\hat{w})),$$

is unique and given by the following minimization function

$$\hat{w} = \delta_M(t) = \phi'(t)/2t.$$

Table 2.2 tabulates five commonly used half-quadratic functions and their corresponding minimization functions. We observe that these functions have similar properties and achieve the minima (zero) at origin. They all belong to M-estimator in Sect. 2.1 and naturally robust to outliers. ℓ_1-ℓ_2 potential function takes both the advantage of absolute function ($|.|$) to reduce the influence of large errors and that of ℓ estimator (t^2) to be continuous. "Fair" potential function defines continuous derivatives of the first three orders and yields a unique solution. "Log-cosh" potential function is a strictly convex function and is an approximation of Huber

Table 2.2 Minimization functions δ relevant to the multiplicative and the additive form of HQ for different potential functions ϕ. α in M-estimator is a constant

	Potential function $\phi(t)$	Multiplicative form $\delta_M(t)$	Additive form $\delta_A(t)$
$\ell_1 - \ell_2$	$\sqrt{\alpha + t^2} - 1$ 	$1/\sqrt{\alpha + t^2}$ 	$t - \frac{t}{\sqrt{\alpha+t^2}}$ 
Fair	$\frac{\lvert t\rvert}{\alpha} - \log(1 + \frac{\lvert t\rvert}{\alpha})$ 	$\frac{1}{\alpha(\alpha+\lvert t\rvert)}$ 	$t - \frac{t}{\alpha(\alpha+\lvert t\rvert)}$ 
log-cosh	$\log(\cosh(\alpha t))$ 	$\alpha \frac{\tanh(\alpha t)}{t}$ 	$t - \alpha \tanh(\alpha t)$ 
Huber	$\begin{cases} t^2/2 & \lvert t\rvert \le \lambda \\ \lambda \lvert t\rvert - \frac{\lambda^2}{2} & \lvert t\rvert > \lambda \end{cases}$ 	$\begin{cases} 1 & \lvert t\rvert \le \lambda \\ \frac{\lambda}{\lvert t\rvert} & \lvert t\rvert > \lambda \end{cases}$ 	$\begin{cases} 0 & \lvert t\rvert \le \lambda \\ t - \lambda\, sign(t) & \lvert t\rvert > \lambda \end{cases}$ 
Welsch	$1 - \exp(-\frac{t^2}{\sigma^2})$ 	$\exp(-\frac{t^2}{\sigma^2})$ 	$t - t\exp(-\frac{t^2}{\sigma^2})$ 

potential function [77], which is a parabola in the vicinity of zero and increases linearly at a given level $|t| > \lambda$. Angst et al. [4] show that Huber function can efficiently handle outliers than ℓ_1 estimator for motion problems. "Welsch" potential function is widely used in information theoretic learning. It has been proved that the robustness of correntropy [99] based algorithms is actually related to Welsch function. In addition, they all can be used as an approximator of ℓ_0-norm to enhance sparsity [69].

The convergency of HQ optimization was justified in [3, 112]. Nikolova and NG [112] further derived the upper bound of the root-convergence for both multiplicative and additive reformulations, showing that the bound of multiplicative form is lower than the additive form. Also, they showed that the number of iterations for multiplicative form is less than additive form, but the computation time of additive form is less. They suggested to use the additive form if possible. Allain et al. [3] also showed that the additive form is faster. They pointed out that these algorithms are special cases of generalized Weiszfeld algorithms [152].

Interestingly, some widely used optimization algorithms, e.g., the robust M-estimator and mean-shift, can be viewed as special cases of HQ optimization. As is well known that HQ remains a local minimization algorithm. Global minimization via HQ can be achieved by applying the graduated non-convexity method [12] for visual reconstruction. Now HQ is widely used in the field of penalized image reconstruction and restoration, such as SAR [20], MRI [132], and spectrometry [105].

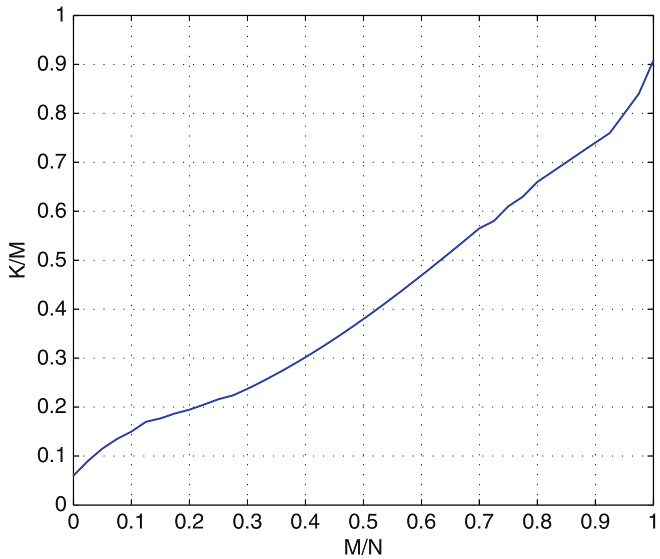


Fig. 2.2 Phase transition diagram corresponding to a compressed sensing or robust system where A is the random Gaussian matrix. The boundary separates regions in the problem space. Below the boundary, signal is well reconstructed; and above the boundary, the system lacks sparsity or robustness, and/or too few measurements are obtained to solve the reconstruction problem correctly [42, 44, 118]

2.3 Phase Transition Diagrams

Phase transition diagrams are often used to evaluate a robust or compressed sensing system [42, 44]. Given a particular system, governed by a sensing matrix A, let $\delta = M/N$ be a normalized measure of under-sampling factor and $\rho = K/M$ be a normalized measure of sparsity or robustness. M, N, and K indicate the number of features, the number of samples, and the number of nonzero entries (or the number of corrupted entries), respectively. A plot of the pairing of the variables δ and ρ describes a 2D phase space $(\delta, \rho) \in [0, 1]$. It has been shown that for many practical sensing matrices, there are sharp boundaries in this phase space. These boundaries clearly divide the solvable problems from unsolvable one in noiseless case. That is, a phase transition diagram provides a way of checking sparsity or robustness, indicating how sparsity and robustness affect the success of a system. Figure 2.2 gives an example of a phase transition diagram which is obtained when sensing matrix A is a random Gaussian matrix. Below the boundary, signal is well reconstructed; and above the boundary, the system lacks sparsity or robustness, and/or too few measurements are obtained to solve the reconstruction problem correctly [118].

2.4 Summary

In this chapter, some preliminary theory and methods related to information theoretic learning have been studied. M-estimation has a long history in robust statistics, which provides a theoretic tool to analyze robustness of information theoretic loss functions. To systematically evaluate algorithmic robustness, phase transition diagram [42, 44] has been introduced, which is a novel and important research trend in compressed sensing. One of merits of information measures is the inclusion of second and higher order information, which also makes these measures difficult to be solved. Many algorithms [59,99,107] have been proposed to simplify the optimization of these measures. Half-quadratic optimization, including the additive and multiplicative forms, has been proved to be an efficient tool to optimize information theoretic measures. One future direction of half-quadratic optimization is developing accelerated algorithms (especially for non-convex loss functions) to save computational costs. Refer to [26, 151] for recent advances of these algorithms.

Springer

http://www.springer.com/978-3-319-07415-3

Robust Recognition via Information Theoretic Learning
He, R.; Hu, B.; Yuan, X.; Wang, L.
2014, XI, 110 p. 29 illus., 25 illus. in color., Softcover
ISBN: 978-3-319-07415-3