Appendix I. Constraint Optimization and Karush-Kuhn-Tucker Condition

Chuwei Zhou

Support Vector Machine

For this composition, the main unfinished musical notes in the five-line staff for the basic theory of support vector machine:

- How to find the optimal with inequality constraints/equality constraints
- Why the minimum of the primal Lagrangian and the maximum of the dual Lagrangian satisfy the strong duality

This part of appendix is to try to explain more about these two points for more mathematically convincing.

Appendix I.a

Constraint Optimization on Equality Constraints

Considering the constrained optimization problem: $\min_{x\in\mathbb{R}^2} f\left(\vec{x}\right) \quad s.t. \quad h\left(\vec{x}\right)=0 \qquad (I.1)$

To decrease the objective function, we should find an infinitesimal differential $\delta\vec{x}$ satisfying

$$f\left(\vec{x}_F+\alpha\delta\vec{\mathrm{x}}\right)< f\left(\vec{x}_F\right) \qquad (I.2)$$

where $\alpha > 0$, and with the Taylor expansion asymptotical to the first order term:

$$f\left(\vec{x}_F+\alpha\delta\vec{\mathrm{x}}\right)= f\left(\vec{x}_F\right)+\alpha\delta\vec{\mathrm{x}}\cdot\nabla_x f\left(\vec{x}_F\right) \qquad (I.3)$$

But the differential $\delta\vec{x}$ should be subjected to the equality constraint $h\left(\vec{x}\right)=0$,

$$h\left(\vec{x}+\alpha\delta\vec{x}\right)=h\left(\vec{x}\right)+\alpha\delta\vec{x}\cdot\nabla h\left(\vec{x}\right)=h\left(\vec{x}\right)=0 \qquad (I.4)$$

which means that $\delta\vec{x}\perp\nabla h\left(\vec{x}\right)$, $\delta\vec{x}=\left\|\delta\vec{x}\right\|\cdot\dfrac{\nabla\times\nabla h\left(\vec{x}\right)}{\left\|\nabla\times\nabla h\left(\vec{x}\right)\right\|}$. Substituting this constraint back to the formula $\left(I.3\right)$, we can obtain

$$f\left(\vec{x}_F+\alpha\delta\vec{\mathrm{x}}\right)= f\left(\vec{x}_F\right)+\frac{\alpha\left\|\delta\vec{x}\right\|}{\left\|\nabla\times\nabla h\left(\vec{x}_F\right)\right\|}\cdot\left[\left(\nabla\times\nabla h\left(\vec{x}_F\right)\right)\cdot\nabla_x f\left(\vec{x}_F\right)\right] \qquad (I.5)$$

Hence when $\nabla h\left(\vec{x}_F\right)//\nabla_x f\left(\vec{x}_F\right)$, $\left(\nabla\times\nabla h\left(\vec{x}_F\right)\right)\cdot\nabla_x f\left(\vec{x}_F\right)=0$, satisfying the translation invariance $f\left(\vec{x}_F+\alpha\delta\vec{\mathrm{x}}\right)= f\left(\vec{x}_F\right)$, and $\vec{x}_F$ would be a local optimal. And the relation $\nabla h\left(\vec{x}_F\right)//\nabla_x f\left(\vec{x}_F\right)$ can be written as

$$\nabla_x f\left(\vec{x}_F\right)=\mu\nabla_x h\left(\vec{x}_F\right) \qquad (I.6)$$

More generally, we construct the so-called primal Lagrangian with the multiplier μ ,

$$L\left(\vec{x},\mu\right)= f\left(\vec{x}\right)-\mu\cdot h\left(\vec{x}\right) \qquad (I.7)$$

Taking $\nabla_x L(\vec{x}, \mu) = 0$ would give us the relation in $(I.6)$, and $\frac{\partial}{\partial \mu}\left[L(\vec{x}, \mu)\right] = 0$ would come up with the equality constraint $h(\vec{x}) = 0$. So, for simplicity, $\nabla_x L(\vec{x}, \mu) = 0$ and $\frac{\partial}{\partial \mu}\left[L(\vec{x}, \mu)\right] = 0$ is the sufficient condition (efficient condition as well) for the local optimal. The remaining is to verify an optimal is maximum or minimum. When taking the Taylor expansion in the formula $(I.3)$, we should improve the accuracy to the second order term,

$$\begin{aligned} f(\vec{x}_F + \alpha\delta\vec{\mathrm{x}}) &= f(\vec{x}_F) + \alpha\delta\vec{\mathrm{x}} \cdot \nabla_x f(\vec{x}_F) + \frac{1}{2}(\alpha\delta\vec{\mathrm{x}} \cdot \nabla_x)^2 f(\vec{x}_F) \\ &= f(\vec{x}_F) + \frac{1}{2}\left(\alpha(\delta\vec{\mathrm{x}})^T \cdot (\nabla_x)^T\right)\left[\alpha\delta\vec{\mathrm{x}} \cdot \nabla_x f(\vec{x}_F)\right] \\ &= f(\vec{x}_F) + \frac{1}{2}\alpha^2 (\delta\vec{\mathrm{x}})^T \cdot \nabla_{xx}^2 f(\vec{x}_F) \cdot (\delta\vec{\mathrm{x}}) \end{aligned} \qquad (I.8)$$

So, when $(\delta\vec{\mathrm{x}})^T \cdot \nabla_{xx}^2 f(\vec{x}_F) \cdot (\delta\vec{\mathrm{x}}) > 0$, $\vec{x}_F$ would be a local minimum. And from the formula $(I.7)$ we know that $L(\vec{x}, \mu)$ and $f(\vec{x})$ would hit the minimum simultaneously. Hence we can conclude the optimality situation for the local minimum of the constrained optimization problem $\min_{x \in \mathbb{R}^n} f(\vec{x}) \quad s.t. \quad h_i(\vec{x}) = 0$ for $i = 1, 2, \dots, l$. is

$$\begin{cases} \nabla_x L(\vec{x}^*, \vec{\mu}^*) = 0 \\ \nabla_\mu L(\vec{x}^*, \vec{\mu}^*) = 0 \\ (\delta\vec{\mathrm{x}})^T \cdot \nabla_{xx}^2 L(\vec{x}^*, \vec{\mu}^*) \cdot (\delta\vec{\mathrm{x}}) > 0 \quad s.t. \quad \delta\vec{x} = \pm\|\delta\vec{x}\| \cdot \dfrac{\nabla \times \nabla h_i(\vec{x})}{\|\nabla \times \nabla h_i(\vec{x})\|} \end{cases} \qquad (I.9)$$

Appendix I.b

Constraint Optimization on Inequality Constraints

Considering the constrained optimization problem: $\min_{x \in \mathbb{R}^2} f(\vec{x}) \quad s.t. \quad g(\vec{x}) \le 0$,

I.b.1. If the unconstrained local minimum of $f(\vec{x})$ is $\vec{x}^*$ satisfying $g(\vec{x}^*) < 0$, which means the local optimal is inside the feasible region, then the corresponding optimality situation for the local minimum is:

$$\begin{cases} \nabla_x f(\vec{x}^*) = 0 \\ g(\vec{x}^*) < 0 \\ (\delta\vec{\mathrm{x}})^T \cdot \nabla_{xx}^2 f(\vec{x}^*) \cdot (\delta\vec{\mathrm{x}}) > 0 \quad \forall \delta\vec{\mathrm{x}} \end{cases} \qquad (I.10)$$

I.b.2. If the unconstrained local minimum $\vec{x}^*$ is not inside the feasible region, it means that we should come up with the minimum point(s) in the boundary of the feasible region $g\left(\vec{x}^*\right)=0$. Hence we can iterating the discussion in $(I.b.1)$ and consider an infinitesimal differential in one point at the boundary $g\left(\vec{x}+\alpha\delta\vec{x}\right)=g\left(\vec{x}\right)=0$, where a slight difference from $(I.b.1)$ should be considered, both $g\left(\vec{x}+\alpha\delta\vec{x}\right)=g\left(\vec{x}\right)+\alpha\delta\vec{x}\cdot\nabla g\left(\vec{x}\right)=0$ and $g\left(\vec{x}+\alpha\delta\vec{x}\right)=g\left(\vec{x}\right)+\alpha\delta\vec{x}\cdot\nabla g\left(\vec{x}\right)\le 0$ can happen in the boundary of the feasible region $g\left(\vec{x}\right)\le 0$. So when we construct the optimal situation for the local minimum based on formula $(I.6)$, we should rephrase the condition like

$$\nabla_x f\left(\vec{x}_F\right)=-\lambda\nabla_x g\left(\vec{x}_F\right) \qquad (I.11)$$

where $\lambda \ge 0$. Because $g\left(\vec{x}+\alpha\delta\vec{x}\right)=g\left(\vec{x}\right)+\alpha\delta\vec{x}\cdot\nabla g\left(\vec{x}\right)\le 0$ gives us $\delta\vec{x}\cdot\nabla g\left(\vec{x}\right)\le 0$, for the $\delta\vec{x}\cdot\nabla g\left(\vec{x}\right)<0$, $f\left(\vec{x}_F+\alpha\delta\vec{\mathrm{x}}\right)=f\left(\vec{x}_F\right)+\alpha\delta\vec{\mathrm{x}}\cdot\nabla_x f\left(\vec{x}_F\right)=$ $f\left(\vec{x}_F\right)-\alpha\lambda\delta\vec{\mathrm{x}}\cdot\nabla_x g\left(\vec{x}_F\right)>f\left(\vec{x}_F\right)$, where guarantees that $\vec{x}_F$ is a local minimum (sufficient condition). Similarly we construct the Lagrangian

$$L\left(\vec{x},\lambda\right)=f\left(\vec{x}\right)+\lambda g\left(\vec{x}\right) \qquad (I.12)$$

$\nabla_x L\left(\vec{x}^*,\lambda^*\right)=0$ would give us the local optimal condition $(I.11)$. And for the case $(I.10)$ with the local minimum inside the feasible region, we can consistently formulate it in $(I.12)$ by taking $\lambda^*=0$ and $\left(\delta\vec{\mathrm{x}}\right)^T\cdot\nabla_{xx}^2 L\left(\vec{x}^*,\lambda^*\right)\cdot\left(\delta\vec{\mathrm{x}}\right)>0$. Hence combining $I.b.1$ and $I.b.2$, we can obtain the optimality conditions for the local minimum(s) of the constrained optimization problem with inequality constraints:

$$L\left(\vec{x},\lambda\right)=f\left(\vec{x}\right)+\sum_{j=1}^{m}\lambda_j g_j\left(\vec{x}\right) \qquad (I.13)$$

$$\begin{cases}\nabla_x L\left(\vec{x}^*,\vec{\lambda}^*\right)=0\\ \lambda_j^* g_j\left(\vec{x}^*\right)=0, \quad j=1,2,\ldots,m\\ \lambda_j^*\ge 0, \quad j=1,2,\ldots,m\\ g_j\left(\vec{x}^*\right)\le 0, \quad j=1,2,\ldots,m\\ \left(\delta\vec{\mathrm{x}}\right)^T\cdot\nabla_{xx}^2 L\left(\vec{x}^*,\vec{\lambda}^*\right)\cdot\left(\delta\vec{\mathrm{x}}\right)>0\end{cases} \qquad (I.14)$$

In our constrained optimization problem with the quadratic objective function

$f\left(\vec{\omega},b\right)=\frac{1}{2}\left\|\vec{\omega}\right\|^2$ and the inequality constraints $g\left(\vec{\omega},b\right)=1-y_i\cdot\left(\vec{x}_i\cdot\vec{\omega}+b\right)\le 0$, similar to the formula $\left(I.13\right)$, we construct the Lagrangian with the multipliers $\mu_i\ge 0$

$$L\left(\left(\vec{\omega},b\right),\vec{\mu}\right)=\frac{1}{2}\left\|\omega\right\|^2-\sum_{i=1}^{l}\mu_i y_i\left(\vec{x}_i\cdot\vec{\omega}+b\right)+\sum_{i=1}^{l}\mu_i \qquad \left(I.15\right)$$

According to the optimality conditions in $\left(I.14\right)$, we obtain

$$\begin{cases}\nabla_{\omega,b}L\left(\left(\vec{\omega}^*,b^*\right),\vec{\mu}^*\right)=0\rightarrow\begin{cases}\vec{\omega}^*-\sum_{i=1}^{l}\mu_i^* y_i\vec{x}_i=0\\ \sum_{i=1}^{l}\mu_i^* y_i=0\end{cases}\\ \mu_i^* g_i\left(\vec{\omega}^*,b^*\right)=\mu_i^*\left[1-y_i\cdot\left(\vec{x}_i\cdot\vec{\omega}+b\right)\right]=0,\quad i=1,2,...,l\\ \mu_i^*\ge 0,\quad i=1,2,...,l\\ g_i\left(\vec{\omega}^*,b^*\right)=1-y_i\cdot\left(\vec{x}_i\cdot\vec{\omega}+b\right)\le 0,\quad i=1,2,...,l\end{cases} \qquad \left(I.16\right)$$

And then with the condition $\vec{\omega}^*-\sum_{i=1}^{l}\mu_i^* y_i\vec{x}_i=0$ and $\sum_{i=1}^{l}\mu_i^* y_i=0$ substituted to our Lagrangian, we can construct the dual Lagrangian：

$$L_D=\sum_{i}^{l}\mu_i-\frac{1}{2}\sum_{i,j}^{l}\mu_i\mu_j y_i y_j\left(\vec{x}_i\cdot\vec{x}_j\right) \qquad \left(I.17\right)$$

Then we take the zero gradient of L_D, $\nabla_{\mu}L_D=0$ giving the a group of l equations to l variables: $1-\sum_{j=1}^{l}\mu_j y_i y_j\left(\vec{x}_i\cdot\vec{x}_j\right)=0$, where $i=1,2,...,l$. Hence, we solve the equations to obtain the local optimal $\vec{\mu}^*$ for the dual Lagrandian. And obviously $\vec{\mu}^*$ is the local maximum for $\max\left\{L_D\right\}$, the maximum lower bound for the $L\left(\left(\vec{\omega},b\right),\vec{\mu}\right)$. And we can easily find that $\min\left\{L\left(\left(\vec{\omega},b\right),\vec{\mu}\right)\right\}\le\min\left\{\frac{1}{2}\left\|\omega\right\|^2\right\}$ from the formula $\left(I.15\right)$ because of $\mu_i\ge 0$, and $g\left(\vec{\omega},b\right)=1-y_i\cdot\left(\vec{x}_i\cdot\vec{\omega}+b\right)\le 0$. And when the condition in formula $\left(I.16\right)$ $\mu_i^* g_i\left(\vec{\omega}^*,b^*\right)=\mu_i^*\left[1-y_i\cdot\left(\vec{x}_i\cdot\vec{\omega}+b\right)\right]=0$ holds for the local optimal, this relation would satisfy the equality $\max\left\{L_D\right\}=\min\left\{L\left(\left(\vec{\omega},b\right),\vec{\mu}\right)\right\}=\min\left\{\frac{1}{2}\left\|\omega\right\|^2\right\}$, the so-called strong duality. And more mathematically, the conditions $\left(I.16\right)$ are the so-called Karush-Kuhn-Tucker conditions. Actually in the formulating framework above, for the sake of simplicity we combine the active inequality constraints and the inactive ones by introducing the non-negative multipliers λ_i, to take the constraints and put them into the objective function, with some λ_k equal to zero for the inactive inequality constraints. This is the so-called complementary slackness, which is part of the more comprehensive Karush-Kuhn-Tucker optimality conditions. The logical foundation supporting this part of Appendix is that for convex problems, the Karush-Kuhn-Tucker optimality conditions are even sufficient and not only necessary. And easily it can be verified that our quadratic optimization problem with linear inequality constraints in the SVM model is a convex optimization problem satisfying the Slater's condition, which is the important premise for that the strong duality and the karush-Kuhn-Tucker conditions hold for the SVM optimization problems. More detail introducing to duality in convex optimization can be referred to the paperwork written by Stephen Wolf (2011).